Human Molecular Genetics, 2005, Vol. 14, No. 9 ***1119–1125***
doi:10.1093/hmg/ddi124
Advance Access published on March 16, 2005

Inferring gene transcriptional modulatory relations: a genetical genomics approach

Hongqiang Li[1,2], Lu Lu[2,3], Kenneth F. Manly[2,3,5], Elissa J. Chesler[3], Lei Bao[1,2], Jintao Wang[2], Mi Zhou[1,2], Robert W. Williams[2,3,4] and Yan Cui[1,2,*]

[1]Department of Molecular Sciences, [2]Center of Genomics and Bioinformatics, [3]Department of Anatomy and Neurobiology, [4]Department of Pediatrics and [5]Department of Pathology and Laboratory Medicine, University of Tennessee Health Science Center, Memphis, TN 38163, USA

Received November 24, 2004; Revised and Accepted March 8, 2005

Bayesian network modeling is a promising approach to define and evaluate gene expression circuits in diverse tissues and cell types under different experimental conditions. The power and practicality of this approach can be improved by restricting the number of potential interactions among genes and by defining causal relations before evaluating posterior probabilities for billions of networks. A newly developed genetical genomics method that combines transcriptome profiling with complex trait analysis now provides strong constraints on network architecture. This method detects those chromosomal intervals responsible for differences in mRNA expression using quantitative trait locus (QTL) mapping. We have developed an efficient Bayesian approach that exploits the genetical genomics method to focus computational effort on the most plausible gene modulatory networks. We exploit a dense marker map for a genetic reference population (GRP) that consists of 32 BXD strains of mice made by intercrossing two progenitor strains—C57BL/6J and DBA/2J. These progenitors differ at ~1.3 million known single nucleotide polymorphisms (SNPs), all of which can be exploited to estimate the probability that a gene contains functional polymorphisms that segregate within the GRP. We constructed 66 candidate networks that include all the candidate modulator genes located in the 209 statistically significant *trans*-acting QTL regions. SNPs that distinguish between the two progenitor strains were used to further winnow the list of candidate modulators. Bayesian network was then used to identify the genetic modulatory relations that best explain the microarray data.

INTRODUCTION

Sequential and contingent changes in gene expression strongly influence the development of organisms and responses to the environment. These dynamic biological programs are executed via complex and still poorly defined networks of interactions among genes, transcripts, proteins and numerous small molecules and cofactors. An adequate definition of these flexible and complex molecular circuits is an essential goal of functional genomics. High-throughput methods, including transcriptome analysis and genome sequencing, have generated huge amounts of data that can be exploited to systematically identify gene modulatory networks.

A recent step forward in this direction involves merging complex trait analysis with transcriptome analysis. This genetical genomics (1) approach treats normal variation in the expression of each gene as a quantitative trait. Quantitative trait locus (QTL) mapping methods are then used to identify the chromosomal intervals that harbor sequence variants (polymorphisms) that produce downstream variations in expression (2–13). This approach is called transcriptome QTL mapping. The major limitation of transcriptome QTL mapping is the difficulty in evaluating candidate genes within QTL intervals that are the ultimate source of variation. A QTL region may contain hundreds of potential polymorphic candidates. Although the strong correlation between DNA variations and gene expression levels indicates that the modulator is located in a particular chromosomal interval, transcriptome QTL mapping cannot identify modulator genes.

*To whom correspondence should be addressed at: Department of Molecular Sciences, University of Tennessee Health Science Center, 858 Madison Avenue, Memphis, TN 38163, USA. Tel: +1 9014483240; Fax: +1 9014487360. Email: ycui2@utmem.edu

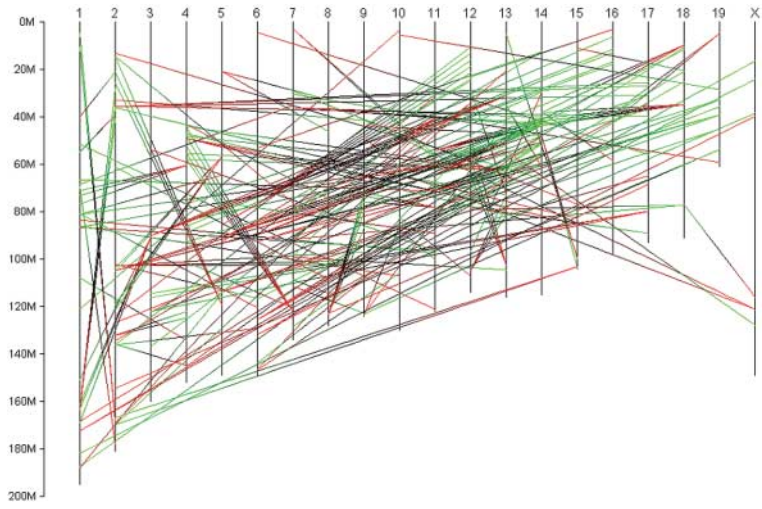

Figure 1. A graphic view of transcriptome QTL mapping results. A total of 175 transcripts have been mapped to 209 *trans*-acting QTLs intervals with LOD scores greater than 4. Mouse chromosomes are represented by vertical lines. The chromosomal locations of genes for all 175 transcripts and their 209 QTLs can be read from the ruler at the leftmost side of the plot. Red/green lines connect a QTL region (represented by a marker) and its target gene. The green end points to the location of the QTL and the red end points to the location of the target gene.

Bayesian network analysis is an effective method to infer the structure of gene regulatory networks from microarray data (4–28). However, constructing putative networks is difficult, because the number of possible networks is a super-exponential function of the number of genes. The total number of possible networks is $3^{[N(N-1)]/2}$, where N is the number of genes. Structure learning of Bayesian networks is therefore an NP-hard problem (29). Here, we show that transcriptome QTL mapping combined with single nucleotide polymorphism (SNP) analysis provides strong constraints on the set of possible upstream modulators of each transcript. Instead of starting with unstructured expression data in which the expression of each transcript can potentially be influenced by any and all other transcripts, we now start from a highly refined set of experimentally supported and directed relations. In this paper, we propose an integrated computational framework based on transcriptome QTL mapping, SNP analysis and Bayesian network. Our method extends beyond mapping of regulatory loci to a systematic evaluation of possible gene modulatory relations using genome-wide genotype, SNP and gene expression data.

The biological motivation of this work is to gain insight into the structure of networks involved in the modulation of gene expression in the mouse brain. All data were obtained from a single genetic reference population (GRP) consisting of 32 BXD recombinant inbred strains. This GRP was generated by crossing two inbred progenitor strains—C57BL/6J and DBA/2J. The genome of each BXD strain is a near-random recombination of chromosome intervals from the two progenitor strains (30,31). Difference in gene expression among members of the GRP can be mapped back to chromosomal intervals using conventional QTL mapping methods. Furthermore, with nearly complete sequence data for both progenitor strains, we can evaluate whether genes within a QTL interval has the type of sequence variants likely to be responsible for a QTL effect.

RESULTS

Transcriptome QTL mapping

The genotypes of the 32 BXD strains have been characterized at several thousand markers, but we used a subset of 779 markers that have been carefully error-checked and that have non-redundant strain distribution patterns. One-hundred Affymetrix U74Av2 arrays were hybridized with pools of mRNA extracted from brain samples of 32 BXD strains, the two parental strains (C57BL/6J and DBA/2J) and their F1 hybrid. Each array was hybridized with mRNA from three animals, and we typically generated three arrays for each strain. For this analysis, we used the Affymetrix MAS 5 transform. Details regarding the experimental conditions, sex and age are available at http://www.genenetwork.org/dbdoc/U74Av2MAS5_December03.html. All of the genotypes and microarray data can be conveniently accessed using WebQTL (9,32,33) (http://www.genenetwork.org). We identified 175 transcripts associated with one or more *trans*-acting QTLs, with a likelihood of odds (LOD) ratio [the relation between the LOD and the likelihood ratio statistics (LRS) is LOD = LRS/4.61] greater than 4.0. These 175 transcripts were mapped to 209 QTL intervals. Figure 1 shows the chromosomal locations of the *trans*-acting QTLs and their downstream target genes. Many significant *cis*-acting QTLs have also been found, but in this work we focus only on the *trans*-acting QTLs to identify the potential upstream modulator genes that may be members of molecular circuits.

Construction of candidate networks

We constructed putative modulatory networks using transcriptome QTL mapping results. Genes located within a QTL region were considered to be candidate sources of variation in downstream mRNA expression. Expression variation in the upstream candidate genes may in turn be mapped to

Table 1. The statistics of the QTL-derived candidate networks

Network no.	Number of QTLs	Number of genes	Number of directed edges
1	72	1395	2397
2	29	387	599
3	6	125	136
4	2	112	210
5	5	91	168
6	8	71	109
7	6	50	106

other QTL intervals. Thus, the target genes and the candidate modulator genes form a network. Each network is a directed graph in which each node represents a gene and each directed edge represents a candidate modulatory relation. We call these networks QTL-derived candidate networks because they contain all the candidate modulatory relations suggested by transcriptome QTL results.

A total of 3123 genes [only the genes in the Affymetrix genechip U74Av2 were counted] are located in the 209 QTL intervals. The transcriptome QTL mapping generated 4815 candidate modulatory relations. We connected the genes of the 175 transcripts with the 3123 genes in the QTL intervals by directed edges representing 4815 candidate modulatory relations. In this way, we constructed 66 QTL-derived candidate networks. Seven of these contained more than 50 genes (Table 1). The largest QTL-derived network contained 1395 nodes (genes) and 2397 directed edges. The 1395 genes were located in 72 QTL intervals scattered on 15 chromosomes (Fig. 2A).

Analysis of the between-strain SNPs

The QTL-derived candidate networks include all the genes in the QTL regions as potential modulator genes. This may lead to very large networks. The complexity of networks can be reduced by eliminating from consideration those genes in QTL intervals that are evidently identical by descent based on the density and distribution of SNPs that distinguish the two progenitor strains. Conversely, candidate genes within QTL intervals that harbored missense and nonsense SNPs were considered very strong candidates. The genomic positions for all RefSeq (34) mRNA transcripts were determined by parsing the corresponding file downloaded from UCSC Genome Browser site (http://genome.ucsc.edu). Curated SNPs between the two mouse strains (~3 million) were also retrieved from Celera SNP database (35). Of these SNPs, ~1.3 million differ between C57BL/6J and DBA/2J. Their genomic positions were determined by BLAT analysis against the mouse genome (36), and missense and nonsense SNPs were screened for each RefSeq mRNA transcript. Genes without missense or nonsense SNPs are less likely to be responsible for effects of the *trans*-acting QTLs, because their protein products are the same in the two progenitor strains and all BXD strains. Such genes and the related edges were removed from the candidate networks. Only 364 genes and 445 candidate modulatory relations survived this process. The resulting networks are called 'QTL-SNP-derived candidate networks'. All 445 candidate modulatory relations are listed in Supplementary Material, Table S1. The biggest QTL-derived candidate network (Fig. 2A) was divided into 15 small networks (Fig. 2B) after filtering by the between-strain missense and nonsense SNPs, with 159 genes and 236 candidate modulatory relations left for further consideration.

Bayesian network modeling

We then used Bayesian network methods to evaluate the sub-networks of the QTL -SNP-derived candidate networks. Under the assumption that there is only one gene in each QTL region modulating the expression of the target gene, we were able to search all the possible network structures exhaustively. Because the Bayesian score is decomposable, we can calculate a score for each target transcript and all candidates independently and select the best scoring modulator(s) for each target transcript. Thus, the total number of scores we need to calculate is

$$N = \sum_{i=1}^{M} \prod_{k=1}^{N_i} n_{ik},$$

where M is the number of target genes in the candidate network; N_i is the number of QTLs associated with the target gene i and n_{ik} is the number of candidate genes in the k th QTL interval of target gene i. We calculated all the possible modulatory relations and predicted 145 modulatory relations that best explained the data (the first 145 modulatory relations in Supplementary Material, Table S1).

Five known transcription factors are involved in six predicted modulatory relations. Three of the five transcription factors have DNA binding matrixes in the TransFac database (37). However, only two predicted target genes of the three transcription factors have annotated 5′-UTRs in their RefSeq (34) sequences, which are needed for retrieving upstream sequences. The 1000 bp upstream regions of the two target genes were extracted from the mouse genome annotation database (http://genome.ucsc.edu). The MATCH program (38) was used to assess whether there was any putative binding site for the predicted modulators in the upstream regions of target genes. The core similarity and the matrix similarity cutoff were set to 1.00 and 0.99, respectively to minimize false positives. The DNA binding motifs were retrieved from the TransFac database. For both target genes, we found DNA binding sites for the predicted modulator in the 1000 bp upstream sequence (Table 2). The core similarity scores (CSS) and matrix similarity scores (MSS) that measure the quality of the match are all equal or very close to 1.0, which denotes perfect match (50).

DISCUSSION

The major challenge in constructing gene modulatory networks from microarray data is that data sets almost invariably contain far fewer samples than needed to specify network architecture. It has been shown that using various constraints can greatly improve the power of Bayesian network (39–41).

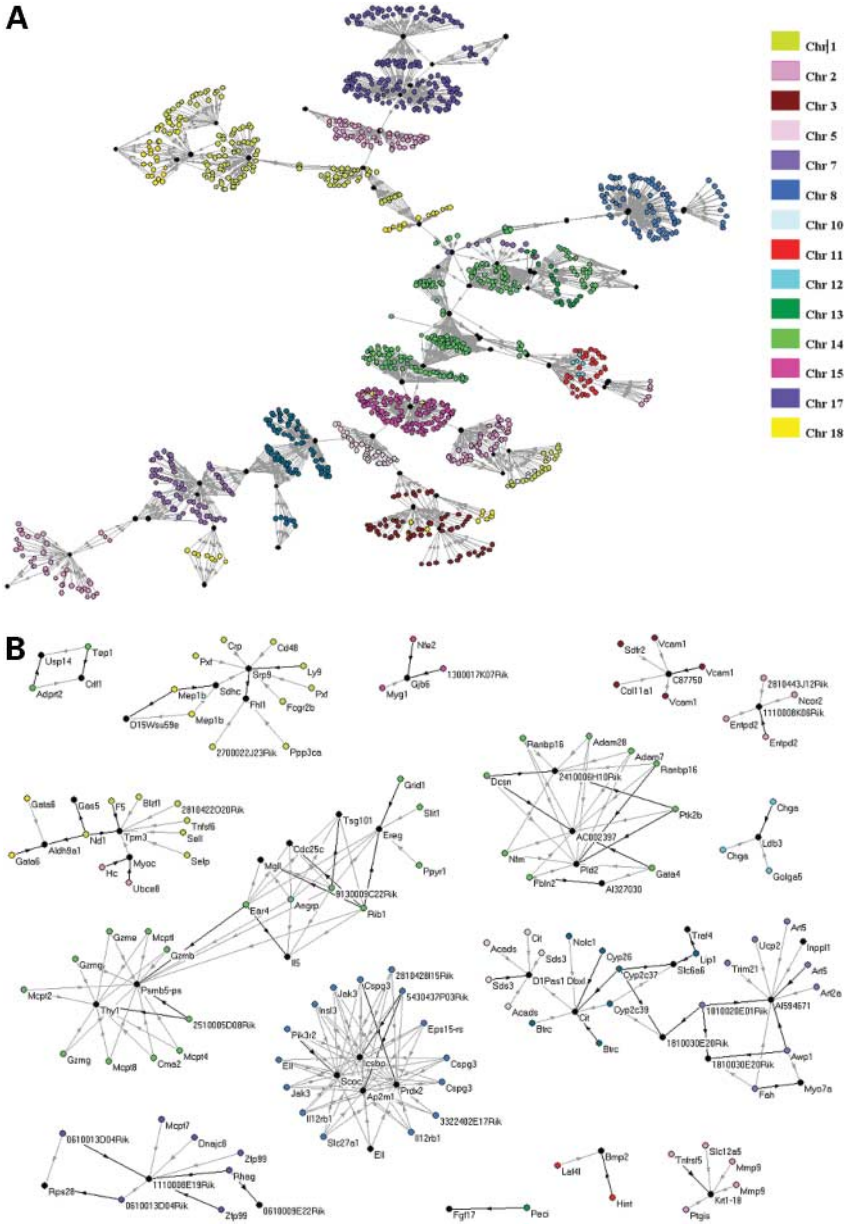

Figure 2. (**A**) A large QTL-derived candidate network that consists of 65 target genes and 1330 candidate modulator genes, all of which are located within 72 QTL intervals on a total of 15 chromosomes. The black nodes represent the target genes. The colored nodes represent potential modulator genes. Colors are used to label chromosomes. The connecting lines (directed edges) represent candidate modulatory relations suggested by the transcriptome-QTL mapping result. (**B**) After filtering by between-strain SNPs, the QTL-derived candidate network is greatly reduced in size and is divided into 15 QTL-SNP-derived candidate networks. The candidate relations are represented by gray and black directed edges. Black edges represent 61 modulatory relations supported by the subsequent Bayesian network analysis. Occasionally, one gene may be represented by more than one node, because it is associated with more than one probe sets on the Affymetrix U74Av2 array.

We addressed this problem by providing a framework for using transcriptome QTL mapping and SNP analysis to dramatically reduce the number of possible network structures that need to be evaluated using subsequent Bayesian network analysis (Fig. 3). Our results provide testable hypotheses on the structures of networks modulating gene expression in the mouse forebrain. Naturally, predicted networks of this type need to be validated or refuted using independent methods and data sets.

Microarray data sets do not in themselves contain sufficient information to reliably construct the whole gene modulatory network. Integrating other data types and sources of information is of crucial importance. In this work, we used genotype data for the BXD GRP in combination with SNP data to cut down the number of gene modulatory relations that need to be evaluated. Other information can also be easily integrated in this framework to provide additional constrains on network structures and to provide suitable prior

Table 2. Examples of predicted modulatory relations with validation from promoter sequence analysis

Modulator gene		Target gene		
Symbol	Gene description	Symbol	Gene description	Binding site for the modulator[a]
Gata6	GATA binding protein 6	Aldh9a1	Aldehyde dehydrogenase 9, subfamily A1	V$GATA6_01 CSS = 1.000 MSS = 0.997
Tcf12	Transcription factor 12	Mela	Melanoma antigen	V$HEB_Q6 CSS = 1.000 MSS = 1.000

[a]The Match program (38) was used to search for the bind sites for transcription factors. Binding site for the predicted modulator was found in the 1000 bp upstream region of the corresponding target gene. Transfac matrix identifiers of the binding sites, CSS and MSS for the matches are shown in this column. The Gata6 binding site is located 374 bp upstream from the transcription start of Aldh9a1, and the Tcf12 binding site is located 177 bp upstream from the transcription start of Mela.

probabilities for the structure learning of the Bayesian network. For example, knowledge encoded in gene function classification systems such as gene ontology (42), MIPS functional catalog (43) and KEGG ontology (44) could potentially be exploited as prior knowledge in the Bayesian network analysis. In principle, the use of multiple types of data enables us to discover the modulatory relations that cannot be inferred from microarray data alone.

In this work, we used the BXD GRP to illustrate the application of a genetical genomics approach. The power of QTL mapping with genetic reference panels will be greatly improved as much larger GRPs are generated (30). The higher the mapping precision, the higher the likelihood that subsequent analysis of candidate modulatory networks will be effective. The framework we describe in this paper can be readily applied to data obtained from *Arabidopsis*, maize, *C aenorhabditis elegans* and *Drosophila*, species for which large RI panels are readily available. The method can also be applied to the whole-genome genotyping and gene expression data obtained from segregating crosses such as F2 intercross and backcross.

Most current molecular networks and pathways are still relatively simple sketches in which many of the key constituents are still missing, misplaced and misdirected. This Bayesian genetic genomics approach allows us to formulate and test larger networks without explicit data on molecular function. It is an efficient method with which to generate new hypotheses that will clearly need to be refuted, verified and refined using additional powerful genetic and molecular methods.

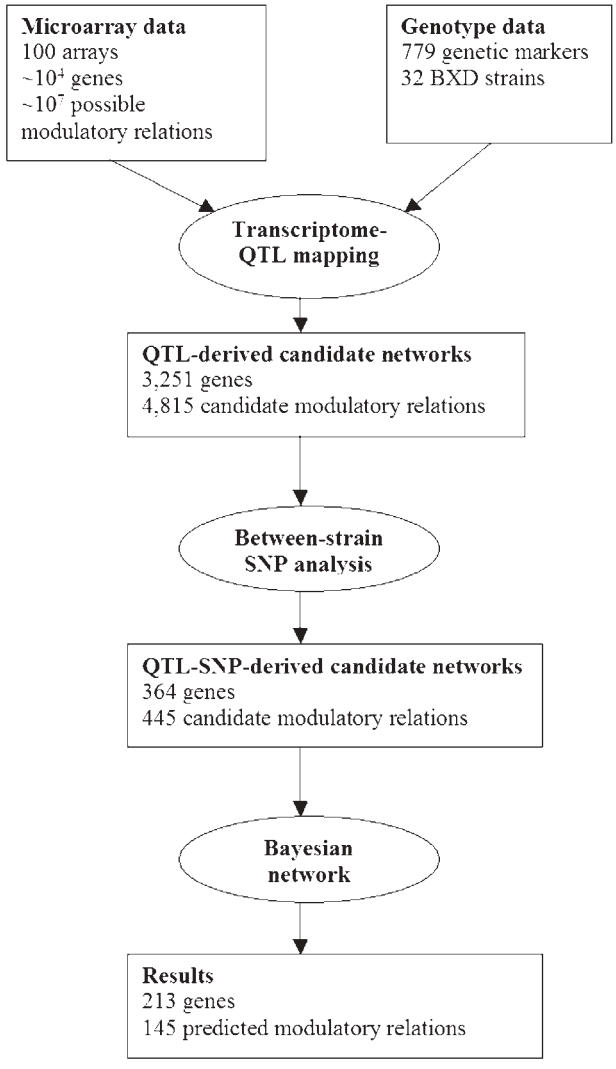

Figure 3. A flow chart of our approach. We begin with two primary data sets: a large gene expression data set generated using 100 Affymetrix U74Av2 arrays and dense marker genotype data for each of 32 members of the BXD GRP. If networks were reconstructed exclusively from microarray data the number of possible modulatory relations would be about 5×10^7 (there are approximately, 10 000 genes represented on the microarray). The transcriptome-QTL mapping results dictate that potential modulator genes must be located within QTL regions, and this reduces the number of relations to 4815. An analysis of between-strain SNPs allows us to further reduce modulatory relations to 445. Finally, the Bayesian network method predicted 145 modulatory relations.

MATERIALS AND METHODS

QTL mapping

The original MAS 5 microarray data were log 2-transformed and normalized to a standard array-wide mean and standard deviation. Values from replicate microarrays were averaged. These values are then evaluated by regression against marker genotypes, where alleles at marker loci were coded as -1 or 1 for the *BB* and *DD* genotypes. The *B* allele is derived from C57BL/6J and the *D* allele is from DBA/2J. Unknown or rare heterozygous markers were coded as 0. The regression model that we used estimates the additive effects of alleles:

$$y_i = b_0 + b_1 x_i + e_i,$$

where y_i, x_i and e_i are the trait value, coded genotype and random environmental effects, respectively, for the *i*th member of the BXD GRP. This allows the regression coefficient to be estimated from sums of squares and sums of products. The least-square estimators for the regression coefficients are

$$\hat{b}_1 = \frac{\sum_{i=1}^{N}(y_i\ x_i) - \left(\sum_{i=1}^{N} y_i \sum_{i=1}^{N} x_i\right)/N}{\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2/N},$$

$$\hat{b}_0 = \bar{y} - \hat{b}_1\bar{x},$$

The LRS was calculated for each regression (37):

$$\mathrm{LRS} = N\ \log\frac{\sum_{i=1}^{N}(y_i - \bar{y})^2}{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2},$$

where N is the number of inbred lines and $\hat{y}_i = b_0 + b_1 x_i$ is the predicted trait value.

This method is the simplest possible approach to QTL mapping. It neglects the possibility of multiple QTLs, dominance and epistatic interactions and it assumes equal variances among GRP strains.

Bayesian network

We applied a Bayesian network method to evaluate subnetworks of the QTL-SNP-derived candidate networks. A Bayesian network (46,47) is a probabilistic graphical model of multiple variables. Given the data set D, one wants to discover the modulatory network that best matches D. The common approach to this problem is to introduce a score to evaluate the posterior probability of a network G given data D:

$$S(G : D) = \log p(G|D),$$

where P is the posterior probability. The Bayesian score for the entire network is decomposable under the assumption of complete data. In the case of a discrete Bayesian network with multinomial local conditional probability distributions, the score can be computed using a closed form equation (39):

$$S(G : D) =$$

$$\log\ P_0 + \sum_{i=1}^{n}\sum_{j=1}^{q_i}\log\left[\frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})}\cdot\prod_{k=1}^{r_i}\frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})}\right],$$

where r_i is the number of states that gene i can assume, q_i denotes the number of joint states that the modulator genes of gene i can have and α_{ijk} is the parameter of Dirichlet prior distribution. (We use a non-informative parameter prior $\alpha_{ijk} = 1/(q_i\ r_i)$ because no prior information about parameters is available (49).) N_{ijk} is the number of occurrences of gene i in state k given parent configuration j, $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$ and $\alpha_{ij} = \sum_{k=1}^{r_i} \alpha_{ijk}$; $\Gamma(\cdot)$ is the gamma function and P_0 is the structure prior. A uniform structure prior was used.

We first normalized the gene expression data for each sample to have the same mean and standard deviation. All expression data were discretized into one of three levels. We calculated the mean (μ) and standard deviation (σ) for each transcripts expression values. If an expression value was less than $\mu - \sigma$ and $\mu + \sigma$, it was assigned to level 0; if an expression value was between $\mu - \sigma$ and $\mu + \sigma$, it was assigned to level 1 and if an expression value was larger than $\mu + \sigma$, it was assigned to level 2.

Network visualization

Kamada–Kawai algorithm (50), a graph layout algorithm implemented in Pajek (51,52) was used to visualize gene networks. Pajek is a program for analyzing and visualizing complex networks.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at HMG Online.

ACKNOWLEDGEMENTS

This work was supported by a PhRMA Foundation grant to Y.C., NIH grants 1U01AA014425-01A1 to L.L., U01AA13499 and MH-62009 to R.W.W. and K.F.M.

REFERENCES

1. Jansen, R.C. and Nap, J.P. (2001) Genetical genomics: the added value from segregation. *Trends Genet.*, **17**, 388–391.
2. Liu, H.C., Cheng, H.H., Tirunagaru, V., Sofer, L. and Burnside, J (2001) A strategy to identify positional candidate genes conferring Marek's disease resistance by integrating DNA microarrays and genetic mapping. *Anim. Genet.*, **32**, 351–359.
3. Brem, R.B., Yvert, G., Clinton, R. and Kruglyak, L. (2002) Genetic dissection of transcriptional regulation in budding yeast. *Science*, **296**, 752–755.
4. Eaves, I.A., Wicker, L.S., Ghandour, G., Lyons, P.A., Peterson, L.B., Todd, J.A. and Glynne, R.J. (2002) Combining mouse congenic strains and microarray gene expression analyses to study a complex trait: the NOD model of type 1 diabetes. *Genome Res.*, **12**, 232–243.
5. Manly, K.F., Wang, J., Shou, S., Qu, Y., Chesler, E., Lu, L., Hsu, H.C., Mountz J.D., Threadgill, D.W. and Williams, R.W. (2002) QTL mapping with microarray expression data. 16th International Mouse Genome Conference, San Antonio, TX.
6. Wayne, M.L. and McIntyre, L.M. (2002) Combining mapping and arraying: an approach to candidate gene identification. *Proc. Natl Acad. Sci. USA*, **99**, 14903–14906.
7. Williams, R.W., Manly, K.F., Shou, S., Chesler, E., Hsu, H.C., Mountz, J.D., Wang, J., Threadgill, D.W. and Lu, L. (2002) Massively parallel complex trait analysis of transcriptional activity in mouse brain. 16th International Mouse Genome Conference, San Antonio, TX.
8. Schadt, E.E., Monks, S.A., Drake, T.A., Lusis, A.J., Che, N., Colinayo, V., Ruff, T.G., Milligan, S.B., Lamb, J.R., Cavet, G. *et al.* (2003) Genetics of gene expression surveyed in maize, mouse and man. *Nature*, **422**, 297–302.
9. Chesler, E.J., Lu, L., Wang, J., Williams, R.W. and Manly, K.F. (2004) WebQTL: rapid exploratory analysis of gene expression and genetic networks for brain and behavior. *Nat. Neurosci.*, **7**, 485–486.
10. Morley, M., Molony, C.M., Weber, T.M., Devlin, J.L., Ewens, K.G., Spielman, R.S. and Cheung, V.G. (2004) Genetic analysis of genome-wide variation in human gene expression. *Nature*, **430**, 743–747.
11. Chesler, E.J., Lu, L., Shou, S., Qu, Y., Gu, J., Wang, J., Hsu, H.C., Mountz, J.D., Baldwin, N.E., Langston, M.A. *et al.* (2005) Complex trait analysis of gene expression reveals polygenic and pleiotropic networks that modulate nervous system function. *Nat. Genet.*, **37**, 233–242.
12. Bystrykh, L., Weersing, E., Dontje, B., Sutton, S., Pletcher, M.T., Wiltshire, T., Su, A.I., Vellenga, E., Wang, J., Manly, K.F. *et al.* (2005) Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'. *Nat. Genet.*, **37**, 225–232.

13. Hubner, N., Wallace, C.A., Zimdahl, H., Petretto, E., Schulz, H., Maciver, F., Mueller, M., Hummel, O., Monti, J., Zidek, V. *et al.* (2005) Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nat. Genet.*, **37**, 243–253.
14. Freidman, N., Linial, M., Nachman, I. and Peer, D. (2000) Using Bayesian networks to analyze expression data. *J. Comput. Biol.*, **7**, 601–620.
15. Peer, D., Regev, A., Elidan, G. and Friedman, N. (2001) Inferring subnetworks from perturbed expression profiles. *Bioinformatics*, **17**, S215–S224.
16. Hartemink, A.J., Gifford, D.K., Jaakkola, T.S. and Young, R.A. (2001) Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. *Pac. Symp. Biocomput.*, 422–433.
17. Spirtes, P., Glymour, C., Scheines, R., Kauffman, S., Aimale, V. and Wimberly, F. (2001) Constructing Bayesian network models of gene expression networks from microarray data. *Proceedings of the Atlantic Symposium on Computational Biology, Genome Information Systems and Technology.*
18. Yoo, C., Thorsson, V. and Cooper, G.F. (2002) Discovery of causal relationships in a gene-regulation pathway from a mixture of experimental and observational DNA microarray data. *Pac. Symp. Biocomput.*, 486–509.
19. Imoto, S., Goto, T. and Miyano, S. (2002) Estimation of genetic networks and functional structures between genes by using Bayesian networks and nonparametric regression. *Pac. Symp. Biocomput.*, 175–186.
20. Imoto, S., Kim, S., Goto, T., Miyano, S., Aburatani, S., Tashiro, K. and Kuhara, S. (2003) Bayesian network and nonparametric heteroscedastic regression for nonlinear modeling of genetic network. *J. Bioinform. Comput. Biol.*, **1**, 231–252.
21. Chu, T., Glymour, C., Scheines, R. and Spirtes, P. (2003) A statistical problem for inference to regulatory structure from associations of gene expression measurement with microarrays. *Bioinformatics*, **19**, 1147–1152.
22. Husmeier, D. (2003) Reverse engineering of genetic networks with Bayesian networks. *Biochem. Soc. Trans.*, **31**, 1516–1518.
23. Tamada, Y., Kim, S., Bannai, H., Imoto, S., Tashiro, K., Kuhara, S. and Miyano, S. (2003) Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection. *Bioinformatics*, **19**, 227–236.
24. Friedman, N. (2003) Probabilistic models for identifying regulation networks. *Bioinformatics*, **19**, II57.
25. Friedman, N. (2004) Inferring cellular networks using probabilistic graphical models. *Science*, **303**, 799–805.
26. Li, Z. and Chan, C. (2004) Inferring pathways and networks with a Bayesian framework. *FASEB J.*, **18**, 746–748.
27. Lu, X., Wang, X., Huang, Y., Hu, W., Gao, G., Li, Y. and Zhang, X., (2004) On some choices in Bayesian network learning for reconstructing regulatory networks. *Proceedings of RECOMB04*, 126–127.
28. Xia, Y., Yu, H., Jansen, R., Seringhaus, M., Baxter, S., Greenbaum, D., Zhao, H. and Gerstein, M. (2004) Analyzing cellular biochemistry in terms of molecular networks. *Annu. Rev. Biochem.*, **73**, 1051–1087.
29. Chickering, D. (1996) Learning Bayesian network is NP-complete. In Fisher, D. and Lenz, H. (eds), *Learning from Data: Artificial Intelligence and Statistics V.* Springer-Verlag, Heidelberg, pp. 121–130.
30. Peirce, J.L., Lu, L., Gu, J., Silver, L.M. and Williams, R.W. (2004) A new set of BXD recombinant inbred lines from advanced intercross populations in mice. *BMC Genet.*, **5**, 7.
31. Plomin, R., McClearn, G.E., Gora-Maslak, G. and Neiderhiser, J.M. (1991) Use of recombinant inbred strains to detect quantitative trait loci associated with behavior. *Behav. Genet.*, **21**, 99–116.
32. Wang, J., Williams, R.W. and Manly, K.F. (2003) WebQTL: web-based complex trait analysis. *Neuroinformatics*, **1**, 299–308.
33. Chesler, E.J., Wang, J., Lu, L., Qu, Y., Manly, K.F. and Williams, R.W. (2003) Genetic correlates of gene expression in recombinant inbred strains. *Neuroinformatics.*, **1**, 343–357.
34. Pruitt, K.D., Tatusova, T. and Maglott, D.R. (2005) NCBI reference sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **33**, D501–D504.
35. Kerlavage, A., Bonazzi, V., di Tommaso, M., Lawrence, C., Li, P., Mayberry, F., Mural, R., Nodell, M., Yandell, M., Zhang, J. and Thomas, P. (2002) The celera discovery system. *Nucleic Acids Res.*, **30**, 129–136.
36. Kent, W.J. (2002) BLAT—the bLAST-like alignment tool, *Genome Res.*, **12**, 656–664.
37. Matys, V., Fricke, E., Geffers, R., Gossling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A.E., Kel-Margoulis, O.V. *et al.* (2003) TRANSFAC®: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.
38. Kel, A.E, Gossling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O.V. and Wingender, E. (2003) MATCH: a tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.*, **31**, 3576–3579.
39. Imoto, S., Higuchi, T., Goto, T., Tashiro, K., Kuhara, S. and Miyano, S. (2004) Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks. *J. Bioinform. Comput. Biol.*, **2**, 77–98.
40. Zhu, J., Lum, P.Y., Lamb, J., GuhaThakurta, D., Edwards, S.W., Thieringer, R., Berger, J.P., Wu, M.S., Thompson, J., Sachs, A.B. and Schadt, E.E. (2004) An integrative genomics approach to the reconstruction of gene networks in segregating populations. *Cytogenet. Genome Res.*, **105**, 363–374.
41. Zou, M. and Conzen, S.D. (2005) A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. *Bioinformatics*, **21**, 71–79.
42. The Gene Ontology Consortium. (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.
43. Mewes, H.W., Frishman, D., Güldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Münsterkoetter, M., Rudd, S. and Weil, B. (2002) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.*, **30**, 31–34.
44. Kanehisa, M. and Goto, S. (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.
45. Haley, C.S. and Knott, S.A. (1992) A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity*, **69**, 315–324.
46. Pearl, J. (1988) *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann Publishers, CA.
47. Pearl, J. (2000) *Causality: Models, Reasoning, and Inference.* Cambridge University Press, Cambridge UK.
48. Heckerman, D. (1999) A tutorial on learning with Bayesian networks. In Jordan, M. (ed.) *Learning in Graphical Models.* MIT Press, Cambridge, MA.
49. Buntine, W. (1991) Theory refinement on Bayesian networks. Proceedings of the Seventh Annual Conference on Uncertainty in Artificial Intelligence, Morgan Kaufman Publishers, CA, pp. 52–60.
50. Kamada, T. and Kawai, S. (1989) An algorithm for drawing general undirected graphs. *Inform. Process. Lett.*, **31**, 7–15.
51. Batagelj, V. and Mrvar, A. (2003) Pajek—analysis and visualization of large networks. In Jünger, M. and Mutzel, P. (eds), *Graph Drawing Software.* Springer, Berlin, pp. 77–103.
52. Batagelj, V. and Mrvar, A. (1998) Pajek—program for large network analysis. *Connections*, **21**, 47–57.